Substantial interobserver variation of thyroid volume and function by visual evaluation of thyroid ^{99m}Tc scintigraphy

Kerstin K. Soelberg[1], Peter Grupe[2], Henrik Boel-Jørgensen[3], Peter H. Jørgensen[4], Søren Fast[1], Viveque E. Nielsen[1], Laszlo Hegedüs[1] & Steen J. Bonnema[1]

ABSTRACT

INTRODUCTION: ^{99m}Tc-pertechnetate scintigraphy is much used in the evaluation of patients with nodular goitre. We investigated the ability of experienced observers to estimate the thyroid 24-h ^{131}I uptake (RAIU) and the thyroid volume by visual evaluation of the scintigram.

MATERIAL AND METHODS: Two endocrinologists and two nuclear medicine specialists visually evaluated thyroid scintigrams from 171 patients with nodular goitre. The variables were assessed in a blinded fashion according to predefined categories and then compared with the true values. The assessments were repeated after four weeks. Kappa (κ_ω) statistics were used.

RESULTS: There was a low probability (range 6-22%) for the observers to assess the thyroid RAIU correctly. The probability of assessing the thyroid volume correctly was in the 14-22% range. Endocrinologists tended to underestimate the thyroid RAIU, mostly in patients with a RAIU > 30%. All observers significantly underestimated the thyroid volume if this was > 80 ml. There was a low interobserver agreement for the thyroid RAIU assessment (κ_ω-values: 0.03-0.43) as well as for the thyroid volume assessment (κ_ω-values: 0.19-0.48). The corresponding κ_ω-values for the intraobserver agreement were 0.34-0.68 and 0.37-0.62, respectively. Nuclear medicine specialists achieved a significantly higher agreement than endocrinologists in their evaluation of both thyroid parameters.

CONCLUSION: Thyroid ^{99m}Tc scintigraphy has poor interobserver agreement and is inaccurate for assessment of quantitative thyroid parameters, even when performed by experienced specialists.

FUNDING: This study was supported by grants from the Danish Agency for Science, Technology and Innovation

TRIAL REGISTRATION: not relevant.

Thyroid scintigraphy is an important tool in the management of thyroid disorders. In addition to the thyroid disease per se, the amount of tracer taken up by the gland correlates inversely with the whole-body iodine pool, and it is also influenced by factors such as iodinated contrast exposure. ^{99m}Tc pertechnetate (^{99m}Tc) is the tracer most commonly used owing to its short half-life and to its cost and availability. Radioiodine (^{131}I) is much used for treatment of non-toxic goitre in some countries [1]. Thyroid imaging is often needed for goitre size estimation since clinical assessment of the goitre volume is notoriously inaccurate [2]. The thyroid ^{131}I uptake (RAIU) can be measured exactly by the administration of a tracer dose of ^{131}I. However, a preliminary decision on whether the patient is eligible for ^{131}I therapy is often based on the appearance of the thyroid ^{99m}Tc scintigram, and the scintigram is thus used semi-quantitatively for assessment of thyroid RAIU and goitre size. Jarløv et al have previously assessed the extent to which clinicians differ in their evaluation of goitre [3, 4]. Regarding the diagnosis of solitary scintigraphically cold thyroid lesions, they found a moderate inter- as well as intraobserver variation; these results were in line with those of studies of diseases in other organs [5]. However, the validity of thyroid ^{99m}Tc scintigraphy for assessment of quantitative thyroid parameters has received little attention. We therefore investigated whether experienced specialists can make valid assessments of the thyroid RAIU and of goitre size based on a visual evaluation of thyroid ^{99m}Tc scintigrams.

MATERIAL AND METHODS

Study population and design

The scintigrams evaluated in this study were obtained from 171 patients who participated in our previous studies [6-9] on recombinant human thyroid-stimulating hormone (TSH) stimulated ^{131}I therapy in patients with a non-toxic nodular goitre. The characteristics of the patients have been described previously [6-9]. All scintigrams were recorded routinely at the initial visits.

Two highly experienced specialists in endocrinology (E1 and E2) and two specialists in nuclear medicine (N1 and N2) participated. None of the four participants were provided with any information about the patients. Based on a visual judgment of a high quality print of the scintigrams, the physicians were asked – blinded with respect to the other observers – to assess a) the thyroid 24-h RAIU, and b) the thyroid volume according to 16 predefined response categories (RAIU: 5% intervals; thyroid

ORIGINAL ARTICLE

1) Department of Endocrinology, Odense University Hospital
2) Department of Nuclear Medicine, Hospital Lillebælt, Vejle
3) Department of Cardiology, Odense University Hospital
4) Department of Nuclear Medicine, Hospital Lillebælt, Vejle

Dan Med J 2014;61(2):A4768

volume: 10 ml intervals for volumes ≤ 100 ml, 20-40 ml intervals for volumes in the range 101-200 ml, and 100 ml intervals for volumes > 200 ml). The true values were available from our previous studies [6-9]. Thus, the thyroid 24-h RAIU was determined after oral administration of a tracer activity of 0.5 MBq ^{131}I, and the thyroid volume was measured by either magnetic resonance imaging (MRI), computed tomography or ultrasound, depending on the size of the goitre and the set-up of the respective study.

TABLE 1

Assessment accuracy of the 24-h RAIU and the thyroid volume for each observer, based on data from the first evaluation.

	Correct assessments, n (%)		Correct assessments ± 1 category[a], n (%)	
Observer	24-h RAIU (N = 171)	thyroid volume (N = 171)	24-h RAIU (N = 171)	thyroid volume (N = 171)
N1[b]	27 (16)	38 (22)	81 (47)	87 (51)
N2[b]	35 (21)	23 (14)	99 (58)	76 (44)
E1[c]	10 (6)	25 (15)	31 (18)	68 (40)
E2[c]	38 (22)	29 (17)	79 (46)	85 (50)

24-h RAIU = thyroid 24-h ^{131}I-uptake.
a) The numbers of correct assessments including the categories immediately adjacent to the true category.
b) Specialist in nuclear medicine.
c) Specialist in endocrinology.

TABLE 2

Observer agreement on the assessment of the 24-h RAIU and thyroid volume. The observers were compared in six different pairs. Weighted κ statistics were used for the analyses of the RAIU and the volume data (a bootstrap of R = 1,000 was chosen), while ordinary κ statistics were used for analysis of the data on isotope distribution.

	24-h RAIU		Thyroid volume	
	κ_ω	95% CI	κ_ω	95% CI
Interobserver variation				
N1 versus N2[a]	0.425	0.355-0.495	0.354	0.282-0.434
E1 versus E2[b]	0.208	0.160-0.260	0.216	0.156-0.275
N1 versus E1	0.048	0.030-0.074	0.195	0.144-0.262
E2 versus N2	0.317	0.266-0.382	0.366	0.299-0.442
N1 versus E2	0.366	0.292-0.444	0.475	0.397-0.551
E1 versus N2	0.029	0.017-0.049	0.250	0.188-0.324
Intraobserver variation				
N1	0.549	0.481-0.616	0.553	0.486-0.631
N2	0.415	0.338-0.488	0.370	0.276-0.454
E1	0.342	0.276-0.405	0.623	0.550-0.687
E2	0.676	0.611-0.736	0.605	0.522-0.664
Interspecialty variation[c]				
N versus E				
1st evaluation	0.157	0.126-0.193	0.262	0.217-0.309
N versus E				
2nd evaluation	0.273	0.228-0.319	0.296	0.238-0.343

24-h RAIU = thyroid 24-h ^{131}I-uptake; CI = confidence interval.
a) Specialist in nuclear medicine
b) Specialist in endocrinology
c) Data from both endocrinologists and both nuclear medicine physicians, respectively, were pooled (n = 342), and a comparison was made for each evaluation.

The survey was performed in two sessions at an interval of approximately four weeks in order to estimate the intraobserver variation. At the second evaluation, the scintigrams were rearranged to minimize the risk of recognition bias. Data from the first evaluation were used to analyse for accuracy of the assessments, while data from the second evaluation were used for determination of the intraobserver variation.

Statistical analyses

The two-sample Wilcoxon rank-sum test and the t-test were used for non-parametric and parametric data, respectively. The χ^2-test was used for categorical variables. Odds ratios were calculated in order to assess the probability of a correct assessment. For calculation of the inter- and intraobserver agreements, the kappa (κ) statistics were used. κ adjusts for the agreement that can be expected by chance alone. The κ coefficient can attain values between –1 and +1. According to the κ-value, the degree of agreement was characterised as poor ($\kappa < 0.00$), slight ($0.00 \le \kappa \le 0.20$), fair ($0.21 \le \kappa \le 0.40$), moderate ($0.41 \le \kappa \le 0.60$), substantial ($0.61 \le \kappa \le 0.80$), or almost perfect ($0.81 \le \kappa \le 1$). Since there were 16 possible categories for the variables in question, we calculated the weighted kappa (κ_ω) [10]. The level of observer agreement was calculated using the bootstrap technique for inferring confidence values [11]. A bootstrap of R = 1,000 was chosen. κ and κ_ω coefficients were compared as described by Gjørup & Jensen [10]. The statistical software used was STATA version 12.1 (STATA Corp LP, Texas, USA). p-values < 0.05 were considered significant.

Trial registration: not relevant.

RESULTS

The thyroid 24-h ^{131}I-uptake: accuracy of observer assessments

Table 1 shows the total number of correct assessments. The number of correct assessments expanded by one level above or below the true category is also shown. All observers had less than 25% correct assessments, while the highest score was 58% accepting a ± one category range. Observer E1 did significantly worse than the other observers in his assessment of the thyroid RAIU ($p < 0.001$).

Figure 1 presents the number of assessments according to each of the 5% interval categories and in relation to the true values. Thus, the "0%" column represents a correct assessment, while the "–20%" represents a RAIU scored four categories below. The endocrinologists (E1 and E2) tended to underestimate the thyroid RAIU by choosing a category, which on average was more than three and one categories, respectively, below the true category. This corresponds to a mean (SD) dis-

tance to the correct category of –16.3 ± 11.1% for E1 and –5.6 ± 11.9% for E2, in contrast to more precise assessments made by N1 (0.5 ± 12.7%) and N2 (0.0 ± 10.5%). The odds for the observers' ability to estimate the thyroid RAIU correctly were 0.188 (N1), 0.257 (N2), 0.062 (E1) and 0.286 (E2). In 67/171 (39%), 65/171 (38%), 17/171 (10%) and 64/171 (37%) of the cases, observer N1, N2, E1 and E2, respectively, assessed the thyroid RAIU correctly (± one category) in both evaluations.

In order to investigate whether the observers' assessments depended on the thyroid RAIU, an arbitrary cut-off level of 30% was chosen. Both endocrinologists were significantly more accurate in their assessment if the true value of the thyroid RAIU was below 30% (E1: $p < 0.001$; E2: $p < 0.002$), while N2 was significantly better in his estimations with a thyroid RAIU above 30% ($p < 0.001$). The accuracy of N1 was unrelated to the thyroid RAIU ($p = 0.591$). Similar results were found by analysing data from the second evaluation, which supports the absence of a learning effect.

The thyroid 24-h ^{131}I-uptake: observer variation and inter-specialty agreement

The four observers were compared in six different pairs (**Table 2**). The κ_ω-value (0.43) for the two specialists in nuclear medicine was significantly higher than the κ_ω-value (0.21) for the endocrinologists ($p < 0.0001$). No complete agreement was reached among the four observers in any patient. The agreement between the two evaluations was determined for each of the four observers (Table 2). For both specialties there was a significant difference between κ_ω-values (0.55 (N1) versus 0.41 (N2), $p = 0.026$; 0.34 (E1) versus 0.68 (E2), $p < 0.0001$). In 105/171 (61%), 114/171 (67%), 134/171 (78%) and 98/171 (57%) of the cases, observer N1, N2, E1 and E2, respectively, altered their assessment from the first to the second evaluation.

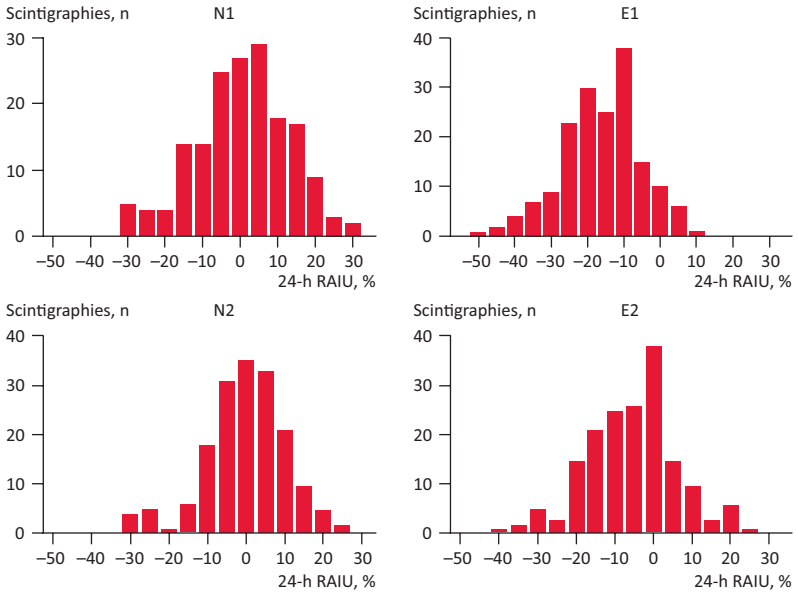
Data from the endocrinologists were pooled, resulting in 342 assessments, as were data from the nuclear medicine physicians. The two specialties were compared for each scintigram, resulting in a κ_ω-value of 0.16 in the first evaluation and of 0.27 in the second evaluation (Table 2).

The thyroid volume: accuracy of observer assessments

Table 1 shows the total number of correct assessments. All observers had less than 25% correct assessments, while the highest score was 51% accepting ± one category. **Figure 2** presents the number of assessments, according to each of the categories, and in relation to the true value, in parallel with the RAIU data. The observers E1 and N2 tended to assess the thyroid volume as being too low, reflected by a deviation from the correct category of mean (SD) –2.84 ± 2.84 for E1 and –1.37 ± 2.99 for N2, in contrast to the more precise assessments made by E2 (–0.91 ± 2.88) and N1 (–0.89 ± 2.90%). The odds for the observers' ability to estimate the thyroid

FIGURE 1

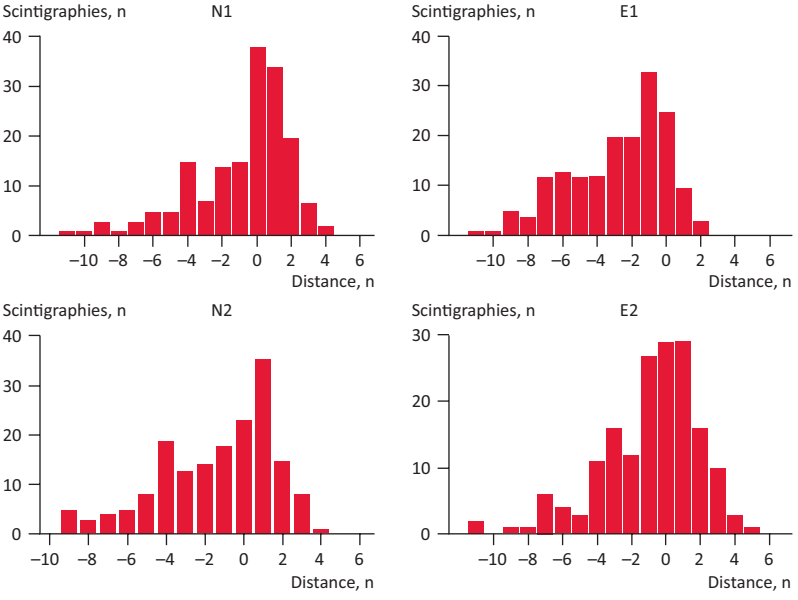
Four histograms, one for each observer, illustrating the number of assessments of the 24-h RAIU according to each category (5% intervals) and in relation to the true category.

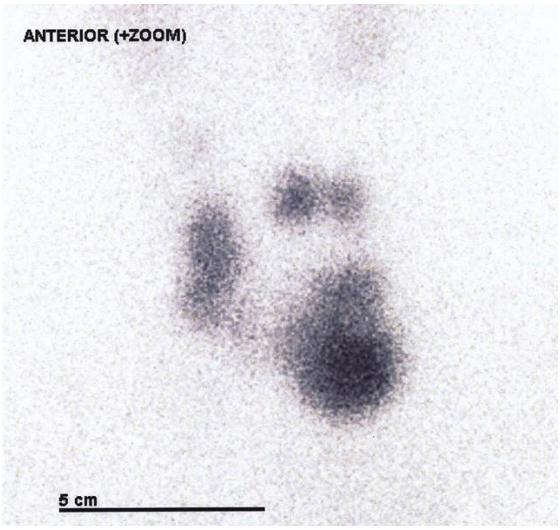


24h RAIU = thyroid 24-hour ^{131}I uptake

FIGURE 2

Four histograms, one for each observer, illustrating the number of assessments of the thyroid volume according to each category and in relation to the true value. Distance is the number of categories from zero (true value).





^{99m}Tc-scintigraphy of a multinodular goitre.

volume correctly were 0.286 (N1), 0.155 (N2), 0.171 (E1) and 0.204 (E2), respectively. In 58/171 (34%), 50/171 (29%), 62/171 (36%) and 61/171 (36%) of the cases, observer N1, N2, E1 and E2, respectively, assessed the thyroid volume correctly (± one category) in both evaluations. If an arbitrary cut-off level of 80 ml was chosen, all four observers were significantly better in estimating thyroid volumes below 80 ml ($p < 0.001$). Similar results were obtained in the repeat evaluation.

The thyroid volume: observer variation and interspecialty agreement

The four observers were compared in six different pairs (Table 2). The κ_ω-value (0.35) for the two specialists in nuclear medicine was significantly higher than the κ_ω-value (0.22) for the endocrinologists ($p = 0.007$). The agreement between the two evaluations was determined for each of the four observers (Table 2). There was a significant difference between κ_ω-values for the two nuclear medicine specialists (0.55 (N1) versus 0.37 (N2), $p = 0.003$), but not for the two endocrinologists. In 114/171 (67%), 116/171 (68%), 91/171 (53%) and 120/171 (70%) of the cases, observer N1, N2, E1 and E2, respectively, altered their assessment from the first to the second evaluation. Similar to the analysis of the RAIU data, the two specialities were compared, which resulted in a κ_ω-value of 0.26 in the first evaluation and of 0.30 in the second evaluation (Table 2).

DISCUSSION

When deciding on the choice of therapy in patients with non-toxic goitre, the size of the gland as well as the thyroid RAIU are crucial parameters [12]. Not all clinicians managing such patients have access to an accurate measurement of the thyroid size or to measures of the thyroid RAIU. At some centres, thyroid scintigraphy is the only examination that supplements the clinical examination and the biochemical tests. A previous study [3] on observers' ability to differentiate between diffuse and multinodular goitres found that a higher agreement was obtained when thyroid scintigraphy was added to other routine tests. However, as demonstrated by the present data, scintigraphic imaging may be misleading in some respects. Thus, the thyroid ^{99m}Tc scintigram is neither a valid method for determination of goitre size nor for determination of thyroid 24-h RAIU. The number of correct 24-h RAIU assessments (within categories of 5% intervals) was low, ranging from 6% to 22%. Even if an assessment was accepted as "correct" by the inclusion of one category below or above the true category, the maximum number of correct assessments only reached 58% for one observer, while the other three scored poorer. The nuclear medicine specialists performed better than their colleagues in endocrinology (particularly due to low performance by one of the endocrinologists), who tended to underestimate the thyroid RAIU, especially in cases with a high thyroid RAIU. As for the assessment of thyroid volume, the accuracy was at a similarly low level. This is in line with the results from previous studies where comparisons between MRI, ultrasound and scintigraphy revealed pronounced differences in thyroid volume estimates [13-15]. In our study, all four observers estimated the thyroid volume significantly more incorrectly when thyroid volumes were above 80 ml. In theory, this may be explained by poorer scintigraphic visualization of large goitres. However, what seems to be a low scintigraphic thyroid uptake is, in fact, a dilution effect of the isotope being distributed into a larger thyroid volume, as supported by our previous study [16]. Indeed, in the present study, we demonstrated that the accuracy of the thyroid volume assessments did not depend on the thyroid RAIU. Cases with such low RAIU (< 10%) were excluded in our study. It can be criticized that the scintigraphy and the measurement of the RAIU were not performed on the same occasion. The thyroid RAIU may have varied, but we reported previously that the RAIU was very stable among our patients in the study period [17].

The κ statistics used in the present study are widely used to quantify the agreement among observers in the evaluation of a specific variable, but κ-values can be misleading when results are compared between studies [18]. The four observers in the present study achieved low to moderate interobserver agreement of their assessments of both the thyroid 24-h RAIU and the thyroid volume, with most κ_ω-values being below 0.41. In none of the 171 scintigrams did all four observers agree on the thyroid RAIU, and only in one case did they agree on thyroid volume. The interobserver agreement was significantly higher for the two specialists in nuclear medi-

cine than for the endocrinologists. Still, high agreement rates do not necessarily imply a high accuracy when compared with a gold standard. In general, we found higher κ_ω-values for the intra- than for the interobserver variation, as commonly seen in such studies [19]. However, inconsistency persists among experienced specialists in their interpretation of a thyroid scintigram, reflected by κ_ω-values in the range of 0.34-0.68 for the intraobserver variation.

In conclusion, the present study underlines that a visual evaluation of thyroid ^{99m}Tc scintigraphy is not useful for a valid assessment of either the thyroid 24-h RAIU or of the thyroid volume. The poor intra- and especially interobserver variation further invalidates the method. Importantly, ^{99m}Tc scintigraphy remains highly useful for differentiating between functioning and non-functioning thyroid nodules, and in these cases it is associated with high observer agreement [20].

CORRESPONDENCE: *Steen J. Bonnema*, Endokrinologisk Afdeling M, Odense Universitetshospital, Søndre Boulevard 29, 5000 Odense C, Denmark. E-mail: steen.bonnema@dadlnet.dk

ACCEPTED: 6 November 2013

CONFLICTS OF INTEREST: Disclosure forms provided by the authors are available with the full text of this article at www.danmedj.dk.

LITERATURE

1. Bonnema SJ, Bennedbæk FN, Wiersinga WM et al. Management of the nontoxic multinodular goitre: a European questionnaire study. Clin Endocrinol (Oxf) 2000;53:5-12.
2. Jarløv AE, Hegedüs L, Gjørup T et al. Accuracy of the clinical assessment of thyroid size. Dan Med Bul 1991;38:87-9.
3. Jarløv AE, Hegedüs L, Gjørup T et al. Observer variation in the clinical assessment of the thyroid gland. J Intern Med 1991;229:159-61.
4. Jarløv AE, Nygaard B, Hegedüs L et al. Observer variation in the clinical and laboratory evaluation of patients with thyroid dysfunction and goiter. Thyroid 1998;8:393-8.
5. Lazarus E, Mainiero MB, Schepps B et al. BI-RADS lexicon for US and mammography: interobserver variability and positive predictive value. Radiology 2006;239:385-91.
6. Bonnema SJ, Nielsen VE, Boel-Jørgensen H et al. Improvement of goiter volume reduction after 0.3 mg recombinant human thyrotropin-stimulated radioiodine therapy in patients with a very large goiter: a double-blinded, randomized trial. J Clin Endocrinol Metab 2007;92:3424-8.
7. Fast S, Hegedüs L, Grupe P et al. Recombinant human thyrotropin-stimulated radioiodine therapy of nodular goiter allows major reduction of the radiation burden with retained efficacy. J Clin Endocrinol Metab 2010;95:3719-25.
8. Nielsen VE, Bonnema SJ, Hegedüs L. Effects of 0.9 mg recombinant human thyrotropin on thyroid size and function in normal subjects: a randomized, double-blind, cross-over trial. J Clin Endocrinol Metab 2004;89:2242-7.
9. Graf H, Fast S, Pacini F et al. Modified-release recombinant human TSH (MRrhTSH) augments the effect of 131I therapy in benign multinodular goiter. Results from a multicenter international, randomized, placebo-controlled study. J Clin Endocrinol Metab 2011;96:1368-76.
10. Gjørup T, Jensen AM. Kappakoefficienten – et mål for reproducerbarhed af nominale og ordinale data. Nord Med 1986;101:90-4.
11. Fung KP, Lee J. Bootstrap estimate of the variance and confidence interval of kappa. Br J Ind Med 1991;48:503-4.
12. Bonnema SJ, Hegedüs L. Radioiodine therapy in benign thyroid diseases: effects, side effects, and factors affecting therapeutic outcome. Endocr Rev 2012;33:920-80.
13. Huysmans DA, de Haas MM, van den Broek WJ et al. Magnetic resonance imaging for volume estimation of large multinodular goitres: a comparison with scintigraphy. Br J Radiol 1994;67:519-23.
14. Wesche MF, Tiel-Van Buul MM, Smits NJ et al. Ultrasonographic versus scintigraphic measurement of thyroid volume in patients referred for 131I therapy. Nucl Med Commun 1998;19:341-6.
15. Bonnema SJ, Andersen PB, Knudsen DU et al. MR imaging of large multinodular goiters: Observer agreement on volume versus observer disagreement on dimensions of the involved trachea. AJR Am J Roentgenol 2002;179:259-66.
16. Bonnema SJ, Fast S, Nielsen VE et al. Serum thyroxine and age rather than thyroid volume and serum TSH - are determinants of the thyroid radioiodine uptake in patients with nodular goiter. J Endocrinol Invest 2011;34:e52-e57.
17. Fast S, Nielsen VE, Grupe P et al. Optimizing 131I uptake after rhTSH stimulation in patients with nontoxic multinodular goiter: Evidence from a prospective, randomized, double-blind study. J Nucl Med 2009;50:732-7.
18. Donker DK, Hasman A, van Geijn HP. Interpretation of low kappa values. Int J Biomed Comput 1993;33:55-64.
19. Koran LM. The reliability of clinical methods, data and judgments (first of two parts). N Engl J Med 1975;293:642-6.
20. Nygaard B, Jarløv AE, Hegedüs L et al. Long-term follow-up of thyroid scintigraphies after 131I therapy of solitary autonomous thyroid nodules. Thyroid 1994;4:167-71.